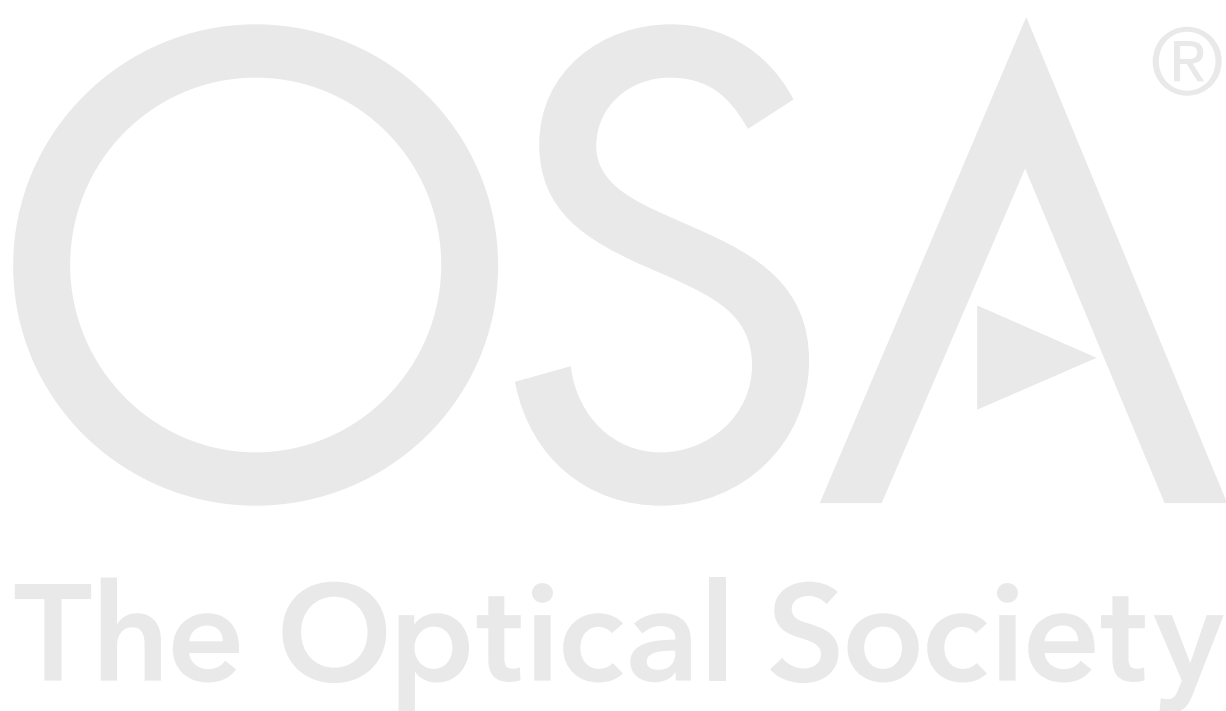


Supplemental document accompanying submission to *Biomedical Optics Express*

Title: Tissue optical properties combined with machine learning enables estimation of articular cartilage composition and functional integrity

Authors: Iman Kafian-Attari, Dmitry Semenov, Ervin Nippolainen, Markku Hauta-Kasari, Juha Töyräs, Isaac Afara

Submitted: 7/16/2020 5:00:46 AM



Supplementary materials

In this section, the impact of each outlier detection technique on the performance of regression models is presented. More specifically, Table S1 presents the performance of the regression models for estimating the PG content and biomechanical properties of articular cartilage when no outlier detection method was utilized.

Table S1. Performance of regression models without any outlier removal in the optical properties, proteoglycan content, and biomechanical properties of articular cartilage on the blind test set. X: No prediction; PLS: partial least-squares regression model; SVR: support vector machines regression model; RF: random forests regression model; Inst.: Instantaneous; Eq.: Equilibrium; and Dyn.: Dynamic modulus. R^2 , ρ , and nRMSE are the coefficient of determination, the correlation coefficient, and the normalized root mean squared by the range of reference variables, respectively. The shaded tabs represent the best estimation of the reference variables.

μ_a	PG_{SZ}	PG_{MZ}	PG_{DZ}	Inst.	Eq.	$Dyn_{-1.0\text{ Hz}}$
Full Spectral range	X	SVR	X	X	SVR	X
1 st NIR window	X	X	X	SVR	X	X
2 nd NIR window	X	X	X	X	X	X
3 rd NIR window	SVR	X	SVR	X	SVR	X
4 th NIR window	X	X	X	SVR	X	X
μ'_S	PG_{SZ}	PG_{MZ}	PG_{DZ}	Inst.	Eq.	$Dyn_{-1.0\text{ Hz}}$
Full Spectral range	X	X	PLS	PLS	PLS	X
1 st NIR window	X	SVR	X	X	PLS	X
2 nd NIR window	X	X	PLS	PLS SVR	X	PLS SVR
3 rd NIR window	X	X	X	PLS SVR	X	X
4 th NIR window	X	X	PLS	PLS	X	X
Model (optimal parameters)	(kernel=rbf, $C=10^2$, $\gamma=2.5 \times 10^{-2}$)	(kernel=sigmoid, $C=10^3$, $\gamma=2.5 \times 10^{-4}$)	($n_{components}=1$)	($n_{components}=1$)	(kernel=rbf, $C=0.25$, $\gamma=0.1$)	($n_{components}=2$)
Score (R^2 , ρ , nRMSE)	(0.0937, 0.4702, 0.1904)	(0.2567, 0.6987, 0.1333)	(0.0721, 0.5583, 0.1705)	(0.1191, 0.7606, 0.1838)	(0.0229, 0.579, 0.2503)	(0.2319, 0.6636, 0.1638)

Table S2 presents the capacity of cartilage optical properties over the aforementioned NIR spectral ranges for estimating the cartilage PG content and biomechanical properties when the boxplot outlier test was employed. It is worth noting that no outlier was detected in the PG content of SZ and equilibrium modulus using this method.

Table S2. Performance of regression models with the boxplot outlier removal in the optical properties, proteoglycan content, and biomechanical properties of articular cartilage on the blind test set. X: No prediction; PLS: partial least-squares regression model; SVR: support vector machines regression model; RF: random forests regression model; split: minimum samples split RF-hyperparameter; Inst.: Instantaneous; Eq.: Equilibrium; and Dyn.: Dynamic modulus. R^2 , ρ , and nRMSE are the coefficient of determination, the correlation coefficient, and the normalized root mean squared by the range of reference variables, respectively. The shaded tabs represent the best estimation of the reference variables.

μ_a	PG_{SZ}	PG_{MZ}	PG_{DZ}	Inst.	Eq.	$Dyn_{1.0\ Hz}$
Full Spectral range	X	X	X	X	SVR	X
1 st NIR window	X	X	X	X	X	X
2 nd NIR window	X	RF	X	X	X	X
3 rd NIR window	SVR	RF	X	X	SVR	X
4 th NIR window	X	X	X	PLS	X	X
μ'_S	PG_{SZ}	PG_{MZ}	PG_{DZ}	Inst.	Eq.	$Dyn_{1.0\ Hz}$
Full Spectral range	X	X	X	X	PLS	X
1 st NIR window	X	X	X	X	PLS	X
2 nd NIR window	X	X	X	X	X	X
3 rd NIR window	X	X	PLS	X	X	X
4 th NIR window	X	X	X	PLS	X	X
Model (optimal parameters)	(kernel=rbf, $C=10^2$, $\gamma = 2.5 \times 10^{-2}$)	($n_{trees}=10$, split=16, no bootstrap)	($n_{components}=1$)	($n_{components}=1$)	(kernel=rbf, $C=0.25$, $\gamma=0.1$)	X
Score (R^2 , ρ , nRMSE)	(0.0937, 0.4702, 0.1904)	(0.0744, 0.6175, 0.2148)	(0.025, 0.5206, 0.2406)	(0.0166, 0.7425, 0.1962)	(0.0229, 0.579, 0.2503)	X

43
44
45
46
47
48
49
50
51
52
53
54
55

Table S3 presents the performance of the regression models when ELEN outlier technique was utilized on the cross-validation and blind test set for detecting the discordant data.

Table S3. Performance of regression models with ELEN outlier removal in the optical properties, proteoglycan content, and biomechanical properties of articular cartilage on the blind test set. X: No prediction; PLS: partial least-squares regression model; SVR: support vector machines regression model; RF: random forests regression model; Inst.: Instantaneous; Eq.: Equilibrium; and Dyn.: Dynamic modulus. R^2 , ρ , and nRMSE are the coefficient of determination, the correlation coefficient, and the normalized root mean squared by the range of reference variables, respectively. The shaded tabs represent the best estimation of the reference variables.

μ_a	PG_{SZ}	PG_{MZ}	PG_{DZ}	Inst.	Eq.	$Dyn_{-1.0\ Hz}$
Full Spectral range	X	X	X	PLS	X	X
				SVR		
1 st NIR window	X	X	X	PLS	X	SVR
				SVR		
2 nd NIR window	X	X	PLS	X	X	X
3 rd NIR window	SVR	X	SVR	X	X	X
4 th NIR window	X	SVR	X	X	X	PLS
μ'_S	PG_{SZ}	PG_{MZ}	PG_{DZ}	Inst.	Eq.	$Dyn_{-1.0\ Hz}$
Full Spectral range	X	SVR	PLS	PLS	X	PLS
			SVR	SVR		
1 st NIR window	X	SVR	X	PLS	X	PLS
2 nd NIR window	X	X	X	PLS	X	PLS
				SVR		
3 rd NIR window	X	X	SVR	PLS	X	PLS
				SVR		
				RF		
4 th NIR window	X	SVR	PLS	PLS	X	X
				RF		
Model (optimal parameters)	(kernel=rbf, $C=10^3$, $\gamma=2.5 \times 10^{-3}$)	(kernel=sigmoid, $C=10^2$, $\gamma=2.5 \times 10^{-4}$)	(kernel=rbf, $C=10^3$, $\gamma=10^{-4}$)	($n_{components}=3$)	X	($n_{components}=2$)
Score (R^2 , ρ , nRMSE)	(0.1835, 0.6387, 0.1367)	(0.565, 0.8267, 0.0638)	(0.0447, 0.7081, 0.158)	(0.4413, 0.9091, 0.1916)	X	(0.294, 0.9167, 0.1978)

56
57
58
59
60
61
62
63
64

Table S4 indicates how the OCSVM outlier detection technique affected the capacity of cartilage optical properties for predicting the composition and functional integrity of the tissue when the regression models developed and tuned on the cross-validation set and evaluated on the blind test set.

Table S4. Performance of regression models with OCSVM outlier removal in the optical properties, proteoglycan content, and biomechanical properties of articular cartilage on the blind test set. X: No prediction; PLS: partial least-squares regression model; SVR: support vector machines regression model; RF: random forests regression model; split: minimum samples split RF-hyperparameter; Inst.: Instantaneous; Eq.: Equilibrium; and Dyn.: Dynamic modulus. R^2 , ρ , and nRMSE are the coefficient of determination, the correlation coefficient, and the normalized root mean squared by the range of reference variables, respectively. The shaded tabs represent the best estimation of the reference variables.

μ_a	PG_{SZ}	PG_{MZ}	PG_{DZ}	Inst.	Eq.	$Dyn_{-1.0\text{ Hz}}$
Full Spectral range	SVR	PLS	PLS	X	X	X
		SVR	SVR			
1 st NIR window	X	X	PLS	X	X	X
2 nd NIR window	X	X	PLS	X	X	X
3 rd NIR window	X	X	X	X	SVR	X
4 th NIR window	SVR	PLS	PLS	PLS	SVR	X
		SVR	SVR			
μ'_S	PG_{SZ}	PG_{MZ}	PG_{DZ}	Inst.	Eq.	$Dyn_{-1.0\text{ Hz}}$
Full Spectral range	X	PLS	PLS	X	X	X
		SVR	RF			
1 st NIR window	X	PLS	X	X	X	X
		SVR				
2 nd NIR window	X	X	X	X	X	X
3 rd NIR window	X	X	X	PLS	X	X
4 th NIR window	X	PLS	PLS	RF	X	X
Model (optimal parameters)	(kernel= sigmoid, C= 10^3 , $\gamma=5 \times 10^{-4}$)	(kernel=rbf, C= 10^3 , $\gamma=2.5 \times 10^{-4}$)	(kernel=sigmoid, C=50, $\gamma=5 \times 10^{-4}$)	($n_{trees}=110$, split=16, bootstrap)	(kernel= sigmoid, C= 10^2 , $\gamma=2.5 \times 10^{-3}$)	X
Score (R^2 , ρ , nRMSE)	(0.035, 0.6574, 0.17)	(0.3959, 0.7682, 0.0875)	(0.2472, 0.6904, 0.1192)	(0.1074, 0.6036, 0.2065)	(0.0737, 0.7253, 0.2661)	X

85 Table S5 presents the capacity of cartilage optical properties over the NIR spectral ranges for
86 estimating the cartilage PG content and biomechanical properties when the LOF outlier
87 method was employed.

88
89 **Table S5. Performance of regression models with LOF outlier removal in the optical properties, proteoglycan**
90 **content, and biomechanical properties of articular cartilage on the blind test set. X: No prediction; PLS:**
91 **partial least-squares regression model; SVR: support vector machines regression model; RF: random forests**
92 **regression model; split: minimum samples split RF-hyperparameter; Inst.: Instantaneous; Eq.: Equilibrium;**
93 **and Dyn.: Dynamic modulus. R^2 , ρ , and nRMSE are the coefficient of determination, the correlation**
94 **coefficient, and the normalized root mean squared by the range of reference variables, respectively. The**
95 **shaded tabs represent the best estimation of the reference variables.**
96

μ_a	PG_{SZ}	PG_{MZ}	PG_{DZ}	Inst.	Eq.	$Dyn_{1.0\text{ Hz}}$
Full Spectral range	SVR	SVR	X	X	X	X
1 st NIR window	X	X	X	X	X	X
2 nd NIR window	X	RF	PLS	X	X	X
3 rd NIR window	SVR	RF	X	X	SVR	X
	RF					
4 th NIR window	PLS	PLS	X	X	X	X
μ'_S	PG_{SZ}	PG_{MZ}	PG_{DZ}	Inst.	Eq.	$Dyn_{1.0\text{ Hz}}$
Full Spectral range	X	X	PLS	PLS	SVR	X
			RF	SVR		
1 st NIR window	X	X	X	X	X	X
2 nd NIR window	X	X	PLS	PLS	X	X
3 rd NIR window	X	X	PLS	PLS	X	SVR
			RF	SVR		
4 th NIR window	X	X	X	PLS	X	SVR
Model (optimal parameters)	(kernel=rbf, C=10, $\gamma=10^{-2}$)	(n _{trees} =10, split=16, bootstrap)	(n _{components} =1)	(n _{components} =3)	(kernel=rbf, C=5, $\gamma=10^{-2}$)	(kernel=sigmoid, C=5, $\gamma=10^{-1}$)
Score (R^2 , ρ , nRMSE)	(0.2335, 0.6397, 0.1984)	(0.1347, 0.5864, 0.2126)	(0.0821, 0.5707, 0.2083)	(0.3461, 0.6904, 0.1535)	(0.0565, 0.6397, 0.2757)	(0.1141, 0.7464, 0.176)

97
98
99
100
101
102
103
104
105
106
107

108 Finally, Table S6 presents the impact of the ISOFOR outlier detection method on the capacity
 109 of cartilage optical properties over the NIR spectral ranges for estimating the cartilage PG
 110 content and biomechanical properties.

111 **Table S6. Performance of regression models with ISOFOR outlier removal in the optical properties,**
 112 **proteoglycan content, and biomechanical properties of articular cartilage on the blind test set. X: No**
 113 **prediction; PLS: partial least-squares regression model; SVR: support vector machines regression model; RF:**
 114 **random forests regression model; split: minimum samples split RF-hyperparameter; Inst.: Instantaneous; Eq.:**
 115 **Equilibrium; and Dyn.: Dynamic modulus. R^2 , ρ , and nRMSE are the coefficient of determination, the**
 116 **correlation coefficient, and the normalized root mean squared by the range of reference variables, respectively.**
 117 **The shaded tabs represent the best estimation of the reference variables.**
 118
 119

μ_a	PG_{SZ}	PG_{MZ}	PG_{DZ}	Inst.	Eq.	$Dyn_{-1.0\text{ Hz}}$
Full Spectral range	PLS	X	X	X	X	X
1 st NIR window	X	X	X	SVR	X	X
2 nd NIR window	X	X	X	X	X	X
3 rd NIR window	SVR	X	SVR	X	SVR	X
4 th NIR window	X	X	X	X	X	X
μ'_S	PG_{SZ}	PG_{MZ}	PG_{DZ}	Inst.	Eq.	$Dyn_{-1.0\text{ Hz}}$
Full Spectral range	X	X	RF	PLS	X	X
				SVR		
1 st NIR window	X	X	X	X	X	X
2 nd NIR window	X	X	RF	X	X	X
3 rd NIR window	X	X	RF	SVR	X	X
				RF		
4 th NIR window	X	X	PLS	RF	X	X
Model (optimal parameters)	($n_{\text{components}}=5$)	X	($n_{\text{trees}}=10$, split=16, bootstrap)	($n_{\text{trees}}=10$, split=16, bootstrap)	(kernel=rbf, $C=10^3$, $\gamma=5 \times 10^{-3}$)	X
Score (R^2 , ρ , nRMSE)	(0.1402, 0.5536, 0.2016)	X	(0.49, 0.7672, 0.1262)	(0.2047, 0.7106, 0.1826)	(0.2213, 0.75, 0.25)	X